

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
13 February 2003 (13.02.2003)

PCT

(10) International Publication Number  
**WO 03/012118 A1**

(51) International Patent Classification<sup>7</sup>: **C12P 19/34**

(21) International Application Number: PCT/US02/23568

(22) International Filing Date: 23 July 2002 (23.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
09/920,491 31 July 2001 (31.07.2001) US

(71) Applicant (*for all designated States except US*):  
**AFFYMETRIX, INC.** [US/US]; 3380 Central Ex-  
pressway, Santa Clara, CA 95051 (US).

(72) Inventor; and

(75) Inventor/Applicant (*for US only*): **DONG, Shoulian**  
[CN/US]; 3350 Mira Vista Circle, San Jose, CA 95132  
(US).

(74) Agents: **LIEBESCHUETZ, Joe** et al.; TOWNSEND  
AND TOWNSEND AND CREW LLP, Two Embarcadero  
Center, Eighth Floor, San Francisco, CA 94111 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU,  
AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,  
CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH,  
GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC,  
LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW,  
MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG,  
SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,  
VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM,  
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),  
Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),  
European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,  
ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK,  
TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,  
GW, ML, MR, NE, SN, TD, TG).

**Declaration under Rule 4.17:**

— *of inventorship (Rule 4.17(iv)) for US only*

**Published:**

— *with international search report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: COMPLEXITY MANAGEMENT OF GENOMIC DNA

(57) Abstract: The presently claimed invention provides for novel methods and kits for reducing the complexity of a nucleic acid sample by providing non-gel based methods for amplification of a subset of the sequences in a sample. In a preferred embodiment, amplification of a subset can be accomplished by digesting a sample with two or more restriction enzymes and ligating adaptors to the fragments so that only a subset of the fragments can be amplified. The invention further provides for analysis of the above amplified sample by hybridization to an array, which may be specifically designed to interrogate the desired fragments for particular characteristics, such as, for example, the presence or absence of a polymorphism.

WO 03/012118 A1

## **COMPLEXITY MANAGEMENT OF GENOMIC DNA**

### **RELATED APPLICATIONS**

**[01]** This application claims the benefit of U.S. Application No. 09/428,350 filed 10/27/99, the entire teachings of which are incorporated herein by reference.

### **FIELD OF THE INVENTION**

**[02]** The invention relates to enrichment and amplification of sequences from a nucleic acid sample. In one embodiment, the invention relates to enrichment and amplification of nucleic acids for the purpose of further analysis. The present invention relates to the fields of molecular biology and genetics.

### **BACKGROUND**

**[03]** The past years have seen a dynamic change in the ability of science to comprehend vast amounts of data. Pioneering technologies such as nucleic acid arrays allow scientists to delve into the world of genetics in far greater detail than ever before. Exploration of genomic DNA has long been a dream of the scientific community. Held within the complex structures of genomic DNA lies the potential to identify, diagnose, or treat diseases like cancer, Alzheimer disease or alcoholism. Exploitation of genomic information from plants and animals may also provide answers to the world's food distribution problems.

**[04]** Recent efforts in the scientific community, such as the publication of the draft sequence of the human genome in February 2001, have changed the dream of genome exploration into a reality. Genome-wide assays, however, must contend with the complexity of genomes; the human genome for example is estimated to have a complexity of  $3 \times 10^9$  base pairs. Novel methods of sample preparation and sample analysis that reduce complexity may provide for the fast and cost effective exploration of complex samples of nucleic acids, particularly genomic DNA.

**[05]** In the future disease discovery and diagnosis will be increasingly done at the genetic level. There are already many diseases that are known to be associated with specific genetic components. That number will rapidly increase as researchers

continue to acquire information about the genetic variation between individuals and to associate genotypic information with phenotypic characteristics.

### SUMMARY OF THE INVENTION

[06] The present invention provides for novel methods of sample preparation and analysis comprising managing or reducing the complexity of a nucleic acid sample. The present invention eliminates the need for multiplex PCR and for many of the embodiments the step of complexity reduction may be performed entirely in a single tube. The invention further provides for analysis of the sample by hybridization to an array which may be specifically designed to interrogate fragments for particular characteristics, such as, for example, the presence or absence of a polymorphism. The invention further provides for methods of using a computer system to model enzymatic reactions in order to determine experimental conditions and to design arrays. In a preferred embodiment the invention discloses novel methods of genome-wide polymorphism discovery and genotyping.

[07] In one embodiment the present invention provides methods for reducing the complexity of a nucleic acid sample by fragmenting the nucleic acid sample using restriction enzymes; ligating adaptors to the fragments; and selectively amplifying the fragments that were cut on one end by one restriction enzyme and on the other end by a different restriction enzyme. The complexity of the resulting sample can be regulated by choosing enzymes that cut with a particular frequency.

[08] In one embodiment of the current invention selective amplification is accomplished by blocking ligation of one strand of each adaptor so that fragments that have the same adaptor ligated to both ends will have gaps in both strands but fragments that have one adaptor ligated to one end and a second adaptor ligated to the other end will have one complete strand for amplification. In this embodiment, ligation is blocked at the 5' end of one adaptor and at the 3' end of the other adaptor.

[09] In another embodiment of the current invention selective amplification is accomplished by designing one adaptor to have a 5' overhang and the second adaptor to have a 3' overhang. The overhangs comprise primer binding sites.

[10] Many of the embodiments of the current invention can further reduce complexity by designing the primers for amplification to bind to one or more nucleotides of the variable region of the restriction fragments.

[11] In yet another embodiment, the invention relates to a kit comprising reagents and instructions for reducing the complexity of a nucleic acid sample. The kit may comprise buffers, restriction enzymes, adaptors, primers, a ligase, a polymerase, dNTPS and instructions necessary for use of the kit.

[12] The present invention also provides methods for genotyping an individual which may further comprise contacting the amplified sequences with a solid support comprising nucleic acid probes, and detecting the presence or absence of hybridization of the amplified sequences to the nucleic acid probes on the solid support. The immobilized probes in a preferred embodiment are capable of interrogating one or more polymorphic sites. The identity of the polymorphic base is determined from the hybridization information.

### BRIEF DESCRIPTION OF THE DRAWINGS

[13] Figure 1 is a schematic illustration of a method of complexity reduction that uses targeted amplification of a subset of restriction fragments wherein the targeting is accomplished by selectively blocking ligation of one strand of each of two adaptors.

[14] Figure 2 is a schematic showing how blocked ligation can be used to selectively amplify the fragments that have been digested by two different restriction enzymes.

[15] Figure 3 shows an example of how an adaptor can be used to introduce a gap, resulting in blocked ligation and blocked extension at the gap.

[16] Figure 4 is a schematic illustration of a method of complexity reduction that uses targeted amplification of a subset of restriction fragments wherein the targeting is accomplished by adding overhangs to the adaptors, which comprise primer binding sites.

### DETAILED DESCRIPTION

#### (A) General

[17] The present invention provides a flexible and scalable method for analyzing complex samples of nucleic acids, such as genomic DNA. These methods are not limited to any particular type of nucleic acid sample: plant, bacterial, animal (including human) total genome DNA, RNA, cDNA and the like may be analyzed using some or all of the methods disclosed in this invention. The word "DNA" may be used below as an example of a nucleic acid. It is understood that this term includes all

nucleic acids, such as DNA and RNA, unless a use below requires a specific type of nucleic acid. This invention provides a powerful tool for analysis of complex nucleic acid samples. From experimental design to isolation of desired fragments and hybridization to an appropriate array, the invention provides for fast, efficient and inexpensive methods of complex nucleic acid analysis.

[18] The present invention relies on many patents, applications and other references for details known to those of the art. Therefore, when a patent, application, or other reference is cited or repeated below, it should be understood that it is incorporated by reference in its entirety for all purposes as well as for the proposition that is recited.

[19] As used in the specification and claims, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "an agent" includes a plurality of agents, including mixtures thereof. An individual is not limited to a human being but may also be other organisms including but not limited to mammals, plants, bacteria, or cells derived from any of the above.

[20] Throughout this disclosure, various aspects of this invention are presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as common individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. The same holds true for ranges in increments of  $10^5$ ,  $10^4$ ,  $10^3$ ,  $10^2$ , 10,  $10^{-1}$ ,  $10^{-2}$ ,  $10^{-3}$ ,  $10^{-4}$ , or  $10^{-5}$ , for example. This applies regardless of the breadth of the range.

[21] The practice of the present invention may employ, unless otherwise indicated, conventional techniques of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example hereinbelow. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques can be found in standard laboratory manuals

such as *Genome Analysis: A Laboratory Manual Series (Vols. I-IV)*, *Using Antibodies: A Laboratory Manual*, *Cells: A Laboratory Manual*, *PCR Primer: A Laboratory Manual*, and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), all of which are herein incorporated in their entirety by reference for all purposes.

[22] Some aspects of the present invention make use of microarrays, also called arrays. Methods and techniques applicable to array synthesis have been described in U.S. Patents Nos. 5,143,854, 5,242,974, 5,252,743, 5,324,633, 5,384,261, 5,424,186, 5,451,683, 5,482,867, 5,491,074, 5,527,681, 5,550,215, 5,571,639, 5,578,832, 5,593,839, 5,599,695, 5,624,711, 5,631,734, 5,795,716, 5,831,070, 5,837,832, 5,856,101, 5,858,659, 5,936,324, 5,968,740, 5,974,164, 5,981,185, 5,981,956, 6,025,601, 6,033,860, 6,040,193, and 6,090,555. All of the above patents incorporated herein by reference in their entireties for all purposes.

#### (B) Definitions

[23] Nucleic acids according to the present invention may include any polymer or oligomer of pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. (See Albert L. Lehninger, *Principles of Biochemistry*, at 793-800 (Worth Pub. 1982) which is herein incorporated by reference in its entirety for all purposes). Indeed, the present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glycosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states.

[24] An "oligonucleotide" or "polynucleotide" is a nucleic acid ranging from at least 2, preferably at least 8, 15 or 20 nucleotides in length, but may be up to 50, 100, 1000, or 5000 nucleotides long or a compound that specifically hybridizes to a polynucleotide. Polynucleotides of the present invention include sequences of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) or mimetics thereof which

may be isolated from natural sources, recombinantly produced or artificially synthesized. A further example of a polynucleotide of the present invention may be a peptide nucleic acid (PNA). (See U.S. Patent No. 6,156,501 which is hereby incorporated by reference in its entirety for all purposes.) The invention also encompasses situations in which there is a nontraditional base pairing such as Hoogsteen base pairing which has been identified in certain tRNA molecules and postulated to exist in a triple helix. "Polynucleotide" and "oligonucleotide" are used interchangeably in this application.

[25] The term "fragment," "segment," or "DNA segment" refers to a portion of a larger DNA polynucleotide or DNA. A polynucleotide, for example, can be broken up, or fragmented into, a plurality of segments. Various methods of fragmenting nucleic acids are well known in the art. These methods may be, for example, either chemical or physical in nature. Chemical fragmentation may include partial degradation with a DNase; partial depurination with acid; the use of restriction enzymes; intron-encoded endonucleases; DNA-based cleavage methods, such as triplex and hybrid formation methods, that rely on the specific hybridization of a nucleic acid segment to localize a cleavage agent to a specific location in the nucleic acid molecule; or other enzymes or compounds which cleave DNA at known or unknown locations. Physical fragmentation methods may involve subjecting the DNA to a high shear rate. High shear rates may be produced, for example, by moving DNA through a chamber or channel with pits or spikes, or forcing the DNA sample through a restricted size flow passage, *e.g.*, an aperture having a cross sectional dimension in the micron or submicron scale. Other physical methods include sonication and nebulization. Combinations of physical and chemical fragmentation methods may likewise be employed such as fragmentation by heat and ion-mediated hydrolysis. See for example, Sambrook *et al.*, "Molecular Cloning: A Laboratory Manual," 3<sup>rd</sup> Ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2001) ("Sambrook *et al.*") which is incorporated herein by reference in its entirety for all purposes. These methods can be optimized to digest a nucleic acid into fragments of a selected size range. Useful size ranges may be from 100, 200, 400, 700 or 1000 to 500, 800, 1500, 2000, 4000 or 10,000 base pairs. However, larger size ranges such as 4000, 10,000 or 20,000 to 10,000, 20,000 or 500,000 base pairs may also be useful.

[26] A number of methods disclosed herein require the use of restriction enzymes to fragment the nucleic acid sample. In general, a restriction enzyme recognizes a specific nucleotide sequence of four to eight nucleotides and cuts the DNA at a site within or a specific distance from the recognition sequence. For example, the restriction enzyme *EcoRI* recognizes the sequence GAATTC and will cut a DNA molecule between the G and the first A. The length of the recognition sequence is roughly proportional to the frequency of occurrence of the site in the genome. A simplistic theoretical estimate is that a six base pair recognition sequence will occur once in every 4096 ( $4^6$ ) base pairs and is considered to be a "rare cutter", while a four base pair recognition sequence will occur once every 256 ( $4^4$ ) base pairs and is considered to be a "frequent cutter". An enzyme with an 8 base pair recognition sequence is also a "rare cutter" cutting once every  $4^8$  base pairs. *In silico* digestions of sequences from the Human Genome Project show that the actual occurrences are even more infrequent, depending on the sequence of the restriction site. Because the restriction sites are rare, the appearance of shorter restriction fragments, for example those less than 1000 base pairs, is much less frequent than the appearance of longer fragments. Many different restriction enzymes are known and appropriate restriction enzymes can be selected for a desired result. (For a description of many restriction enzymes, see New England BioLabs Catalog which is herein incorporated by reference in its entirety for all purposes).

[27] "Adaptor sequences" or "adaptors" are generally oligonucleotides of at least 5, 10, or 15 bases and preferably no more than 50 or 60 bases in length, however, they may be even longer, up to 100 or 200 bases. Adaptor sequences may be synthesized using any methods known to those of skill in the art. For the purposes of this invention they may, as options, comprise templates for PCR primers, restriction sites and promoters. The adaptor may be entirely or substantially double stranded. The adaptor may be phosphorylated or unphosphorylated on one or both strands. Adaptors are particularly useful in one embodiment of the current invention if they comprise short single stranded regions, which are complementary to the single stranded region, also known as an overhang, created by digestion with a restriction enzyme. For example, when DNA is digested with the restriction enzyme *EcoRI* the resulting double stranded fragments are flanked at either end by the single stranded overhang 5'-AATT-3', an adaptor that carries a single stranded overhang 5'-AATT-3' will hybridize to the fragment through complementarity between the overhanging



regions. This “sticky end” hybridization of the adaptor to the fragment facilitates ligation of the adaptor to the fragment, although blunt ended ligation is also possible. Different restriction enzymes cleave DNA to produce different overhangs and adaptors can be designed for any overhang. For a list of many restriction enzymes, their recognition sites and the resulting cleavage products, *see*, for example, New England Biolabs Catalog, which is herein incorporated by reference in its entirety for all purposes, available at [www.neb.com](http://www.neb.com), last visited 7/25/01.

[28] Digestion with two or more enzymes can be used to selectively ligate separate adaptors to either end of a restriction fragment. Restriction fragments that have been generated by two different restriction enzymes will often have two distinct overhangs. A pair of adaptors can be designed to ligate specifically to one or the other overhang so that the fragment will ligate to one adaptor on one end and to the second adaptor on the other end. For example, if a fragment is the result of digestion with *EcoRI* at one end and *BamHI* at the other end, the overhangs will be 5'-AATT-3' and 5'-GATC-3', respectively. An adaptor with an overhang of AATT will be preferentially ligated to one end while an adaptor with an overhang of GATC will be preferentially ligated to the second end.

[29] A single adaptor can also be ligated to both ends of a fragment resulting from digestion with two different enzymes. For example, if the method of digestion generates blunt ended fragments, the same adaptor sequence can be ligated to both ends. Alternatively some pairs of enzymes leave identical overhanging sequences. For example, *BglII* recognizes the sequence 5'-AGATCT-3', cutting after the first A, and *BamHI* recognizes the sequence 5'-GGATCC-3', cutting after the first G; both leave an overhang of 5'-GATC-3'. A single adaptor with an overhang of 5'-GATC-3' may be ligated to both digestion products. Adaptors can also be designed to recreate the original restriction site, create a new restriction site or to eliminate the original site.

[30] Methods of ligation will be known to those of skill in the art and are described, for example in Sambrook *et al.* and the New England BioLabs catalog, both of which are incorporated herein in their entireties by reference for all purposes. Methods include using T4 DNA ligase, which catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in duplex DNA or RNA with blunt or and sticky ends; *Taq* DNA ligase, which catalyzes the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini of two

adjacent oligonucleotides that are hybridized to a complementary target DNA; *E.coli* DNA ligase, which catalyzes the formation of a phosphodiester bond between juxtaposed 5' -phosphate and 3' -hydroxyl termini in duplex DNA containing cohesive ends; and T4 RNA ligase which catalyzes ligation of a 5' phosphoryl-terminated nucleic acid donor to a 3' hydroxyl-terminated nucleic acid acceptor through the formation of a 3' →5' phosphodiester bond, substrates include single-stranded RNA and DNA as well as dinucleoside pyrophosphates; or any other methods described in the art.

[31] "Genome" designates or denotes the complete, single-copy set of genetic instructions for an organism as coded into the DNA of the organism. A genome may be multi-chromosomal such that the DNA is distributed among a plurality of individual chromosomes. For example, in human there are 22 pairs of chromosomes plus a gender associated XX or XY pair.

[32] The term "chromosome" refers to the heredity-bearing gene carrier of a living cell which is derived from chromatin and which comprises DNA and protein components (especially histones). The conventional internationally recognized individual human genome chromosome numbering system is employed herein. The size of an individual chromosome can vary from one type to another with a given multi-chromosomal genome and from one genome to another. In the case of the human genome, the entire DNA mass of a given chromosome is usually greater than about 100,000,000 bp. For example, the size of the entire human genome is about  $3 \times 10^9$  bp. The largest chromosome, chromosome no. 1, contains about  $2.4 \times 10^8$  bp while the smallest chromosome, chromosome no. 22, contains about  $5.3 \times 10^7$  bp.

[33] A "chromosomal region" is a portion of a chromosome. The actual physical size or extent of any individual chromosomal region can vary greatly. The term "region" is not necessarily definitive of a particular one or more genes because a region need not take into specific account the particular coding segments (exons) of an individual gene.

[34] The term subset or representative subset refers to a fraction of a genome. The subset may be, for example, 0.01, 0.1, 1, 3, 5, 10, 25, 50 or 75% of the genome. The partitioning of fragments into subsets may be done according to a variety of physical characteristics of individual fragments. For example, fragments may be divided into subsets according to size, according to the particular combination of restriction sites at

the ends of the fragment, or based on the presence or absence of one or more particular sequences.

[35] An "array" comprises a support, preferably solid, with nucleic acid probes attached to the support. Arrays typically comprise a plurality of different nucleic acid probes that are coupled to a surface of a substrate in different, known locations. These arrays, also described as "microarrays" or colloquially "chips" have been generally described in the art, for example, U.S. Pat. Nos. 5,143,854, 5,445,934, 5,744,305, 5,677,195, 5,800,992, 6,040,193, 5,424,186 and Fodor *et al.*, *Science*, 251:767-777 (1991). Each of which is incorporated by reference in its entirety for all purposes.

[36] Arrays may generally be produced using a variety of techniques, such as mechanical synthesis methods or light directed synthesis methods that incorporate a combination of photolithographic methods and solid phase synthesis methods. Techniques for the synthesis of these arrays using mechanical synthesis methods are described in, *e.g.*, U.S. Pat. No. 5,384,261, and 6,040,193, which are incorporated herein by reference in their entirety for all purposes. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may be nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate. (See U.S. Patent Nos. 5,770,358, 5,789,162, 5,708,153, 6,040,193 and 5,800,992, which are hereby incorporated by reference in their entirety for all purposes.)

[37] Arrays may be packaged in such a manner as to allow for diagnostic use or can be an all-inclusive device; *e.g.*, U.S. Patent Nos. 5,856,174 and 5,922,591 incorporated in their entirety by reference for all purposes.

[38] Preferred arrays are commercially available from Affymetrix under the brand name GeneChip® and are directed to a variety of purposes, including genotyping and gene expression monitoring for a variety of eukaryotic and prokaryotic species. (See Affymetrix Inc., Santa Clara and their website at [www.affymetrix.com](http://www.affymetrix.com).)

[39] Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at a frequency of preferably greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphism may comprise one or more base changes, an insertion, a repeat, or a deletion. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length

polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The first identified allelic form is arbitrarily designated as the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms. A polymorphism between two nucleic acids can occur naturally, or be caused by exposure to or contact with chemicals, enzymes, or other agents, or exposure to agents that cause damage to nucleic acids, for example, ultraviolet radiation, mutagens or carcinogens.

**[40]** Single nucleotide polymorphisms (SNPs) are positions at which two alternative bases occur at appreciable frequency ( $>1\%$ ) in the human population, and are the most common type of human genetic variation. The site is usually preceded by and followed by highly conserved sequences of the allele (*e.g.*, sequences that vary in less than 1/100 or 1/1000 members of the populations). It is estimated that there are as many as  $3 \times 10^6$  SNPs in the human genome.

**[41]** A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the polymorphic site. A transition is the replacement of one purine by another purine or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine or vice versa. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele.

**[42]** The term genotyping refers to the determination of the genetic information an individual carries at one or more positions in the genome. For example, genotyping may comprise the determination of which allele or alleles an individual carries for a single polymorphism or the determination of which allele or alleles an individual carries for a plurality of polymorphisms.

**[43]** Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen *et al.*, *Science* 254, 1497-1500 (1991), and other nucleic acid analogs and nucleic acid mimetics. See US Patent Application No. 6,156,501.

[44] Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations. For stringent conditions, *see*, for example, Sambrook *et al.*, which is hereby incorporated by reference in its entirety for all purposes above.

[45] An individual is not limited to a human being, but may also include other organisms including but not limited to mammals, plants, bacteria or cells derived from any of the above.

[46] *In silico* digestion is a computer aided simulation of enzymatic digests accomplished by searching a sequence for restriction sites. *In silico* digestion provides for the use of a computer system to model enzymatic reactions in order to determine experimental conditions before conducting any actual experiments. An example of an experiment would be to model digestion of the human genome with specific restriction enzymes to predict the sizes of the resulting restriction fragments.

#### (C.) The Process

[47] The present invention provides for novel methods of sample preparation and analysis involving managing or reducing the complexity of a nucleic acid sample, such as genomic DNA, by amplifying a representative subset of the sample. The invention further provides for analysis of the above subset by hybridization to an array, which may be specifically designed to interrogate the desired fragments for particular characteristics, such as, for example, the presence or absence of a polymorphism. The invention is particularly useful when combined with other methods of genome analysis. As an example, the present techniques are useful to genotype individuals after polymorphisms have been identified.

[48] One method that has been used to isolate a subset of a genome is to separate fragments according to size by electrophoresis in a gel matrix. The region of the gel containing fragments in the desired size range is then excised and the fragments are purified away from the gel matrix. The SNP consortium (TSC) adopted this approach in their efforts to discover single nucleotide polymorphisms (SNPs) in the human genome. *See*, Altshuler *et al.*, *Science* 407: 513-516 (2000) and The International

SNP Map Working Group, *Nature* 409: 928-933 (2001) both of which are herein incorporated by reference in their entirety for all purposes.

[49] The present invention provides methods of complexity management of nucleic acid samples, such as genomic DNA, that can be used as an alternative to separation of fragments by gel electrophoresis and purification of fragments from a gel matrix. Generally, the embodiments include the steps of: fragmenting the nucleic acid by digestion with two or more restriction enzymes; ligating adaptors to the ends of the fragments; and selectively amplifying a subset of the fragments that have a first adaptor ligated to one end and a second, different adaptor ligated to the other end. In a preferred embodiment the amplified sequences are then exposed to an array which may or may not have been specifically designed and manufactured to interrogate the amplified sequences. Design of both the complexity management steps and the arrays may be aided by computer modeling techniques. Generally, the steps of the present invention involve reducing the complexity of a nucleic acid sample using the disclosed techniques alone or in combination.

[50] When interrogating genomes it is often useful to first reduce the complexity of the sample and analyze one or more subsets of the genome. Subsets can be defined by many characteristics of the fragments. In a preferred embodiment of the current invention the subsets are defined by the restriction enzymes used to generate the fragments and the resulting structure of the ends of the fragments. In some embodiments, the subset is further defined by the identity of one or more nucleotides flanking one or both of the restriction enzyme sites used to generate the fragment.

[51] The genomic DNA sample of the current invention may be isolated according to methods known in the art, such as PCR, reverse transcription, and the like. It may be obtained from any biological or environmental source, including plant, animal (including human), bacteria, fungi or algae. Any suitable biological sample can be used for assay of genomic DNA. Convenient suitable samples include whole blood, tissue, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair.

[52] In a preferred embodiment of the invention, adaptors are ligated to the ends of the fragments and the fragments are amplified by PCR using primers that are designed to hybridize to sequences in the adaptors. In a particularly preferred embodiment, a single primer pair can be used for amplification.

[53] Figure 1 depicts a schematic illustration of the general steps of a preferred embodiment of the currently claimed invention. First, a sample of genomic DNA is

fragmented with a first restriction enzyme, RE1, which in a preferred embodiment is a rare cutter, and a second restriction enzyme, RE2, which in a preferred embodiment is a frequent cutter, to produce three types of fragments. The most common fragments are those which have been cut on both ends with the frequent base cutter. Those fragments, which have been cut on one end with the frequent base cutter and on one end with the rare base cutter will be less frequent. The least common fragments are those which have been cut on both ends with the rare base cutter. One of the adaptors, A1, is designed to ligate to the overhang left by the rare cutter and the other adapter, A2, is designed to ligate to the overhang left by the frequent cutter. Fragments that are cut twice by RE1 will ligate to adaptor A1 on both ends. Fragments that are cut twice by RE2 will ligate to adaptor A2 on both ends. Those fragments that have been cut once with RE1 and once with RE2 will ligate to adaptor A1 on one end and to adaptor A2 on the other end. The adaptors further comprise a primer binding site.

[54] In the embodiment of the invention illustrated in Figure 1, ligation of one strand of each adapter is blocked. Blocking ligation of one strand of the adaptor to the fragments results in a gap or nick between that strand and one strand of the fragment. The gap or nick will block primer extension. In the current invention, one adaptor is designed so that ligation is blocked between the 3' end of one strand of the adaptor and the 5' end of the fragment and the other adaptor is designed to block ligation between the 5' end of one strand of the adaptor and the 3' end of the fragment.

[55] After adaptor ligation the sample is amplified by PCR. Primer P1, specific for adaptor A1, and primer P2, specific for adapter A2, are added, and the mixture is used as template for PCR. Fragments with two identical adaptors are not amplified because primer extension is blocked in both strands at the introduced gaps or nicks (see, Figure 2) so the first strand of DNA synthesized terminates without the primer binding site needed for second strand synthesis. However, fragments with both adaptors A1 and A2 will be amplified because both introduced gaps or nicks are in the same strand leaving a complete strand to serve as a template for PCR.

[56] In this embodiment, the gaps or nicks may be introduced, for example, by blocking ligation of one strand of the adaptor to the fragment. Both the adaptors and the fragments are double stranded so ligation of the adaptor to the fragment typically results in the formation of two phosphodiester bonds, one between one adaptor strand

and one strand of the fragment and a second between the other adaptor strand and the other strand of the fragment. In one embodiment of the current invention, formation of one of these phosphodiester bonds is blocked. The adaptor is ligated to the fragment through one phosphodiester bond but one strand of the resulting ligated fragment will have a gap between the adaptor and the restriction fragment. When this strand is used as a template for primer extension the primer extension reaction will stop when the nick or gap is reached as illustrated in Figure 2.

[57] Ligation of one strand of the adaptor can be selectively blocked by any available method. In a preferred embodiment of the current invention, the adaptors carry a terminal modification that blocks ligation. For example, the presence of a ddNTP at the 3' end of an adaptor strand prevents ligation because there is no 3' hydroxyl available for the ligation reaction. Other 3' modifications that can be used to block ligation include a 3' phosphate group, phosphorothioate, biotin, fluorescent dye, 2',3' cyclic phosphate, or any other modification that results in the absence of a free 3' hydroxyl. Similarly, the absence of a free phosphate at the 5' end of an adaptor strand will block ligation. Typically oligonucleotides are synthesized without a 5' phosphate, but a phosphate can be added by, for example, a kinasing reaction or be removed by, for example, a phosphatase reaction..

[58] In another preferred embodiment, the adaptors are designed to introduce a gap that blocks ligation. (See, Figure 3). A gap of one or more bases can be introduced between one strand of the adaptor and one strand of the fragment. One method of introducing such a gap is by designing the adaptor overhang with either one extra base or with one fewer base. (See, Figure 3). In the example in figure 3, the adapter designed to hybridize to the overhang left by *Sau3A* is designed to introduce a gap between the 5' phosphate of the adaptor and the 3'OH of the fragment (5'-ATC-3' instead of 5'-GATC-3'). Alternatively, the adaptor designed to hybridize to the overhang left by *EcoRI* in figure 3 will introduce a gap between the 5' phosphate of the fragment and the 3' OH of the adaptor by adding an extra nucleotide to the overhang, (5'-AATTN-3' instead of 5'-AATT-3' where N is any nucleotide).

[59] A combination of methods may also be used, for example, ligation of one adaptor may be blocked by the absence of the necessary phosphate while ligation of the other adaptor is blocked by introduction of a gap of one or more nucleotides.

[60] In this embodiment of the current invention, ligation is preferably blocked at the 5' end of one of the two adaptors and at the 3' end of the other adaptor. More



specifically, if ligation of the first adaptor is blocked between the 3' end of the adaptor and the 5' end of the fragment, then ligation of the second adaptor should be blocked between the 5' end of the adaptor and the 3' end of the fragment. Likewise, if ligation of the first adaptor is blocked between the 5' end of the adaptor and the 3' end of the fragment, then ligation of the second adaptor should be blocked between the 3' end of the adaptor and the 5' end of the fragment. As illustrated in Figure 2, this results in blocked amplification of the fragments with two similar ends but allows for amplification of the fragments with different ends because only those fragments will have a complete template strand.

[61] In another embodiment of the current invention, illustrated in Figure 4, the adaptors are designed with single stranded regions that allow for selective amplification. In the example illustrated in Figure 4, adaptor A1, which is designed to ligate to the overhang left by RE1, has an additional single stranded 3' overhang that serves as a primer binding site for primer P1. Adaptor A2, which is designed to be ligated to the overhang left by RE2, has an additional single stranded 5' overhang that serves as a template for a primer binding site for primer P2. Only fragments that have A1 ligated to one end and A2 ligated to the other end will be amplified. (See Figure 4).

[62] Using the current invention different combinations of enzymes can be used for digestion based upon the predicted frequency of cutting and the desired level of complexity reduction. For example, using a simplistic theoretical calculation, if a 6 base cutter is combined with a four base cutter the resulting complexity reduction is 8 fold. The 6 base cutter is predicted to cut the genome into fragments that are on average  $4^6$  base pairs and the 4 base cutter is predicted to cut the genome into fragments that are on average  $4^4$  base pairs. In the double digest, the  $4^6$  base pair fragments should be digested into  $16 \cdot 4^4$  base pair fragments. Two of these fragments, or  $1/8^{\text{th}}$  of the genome, should have been cut by the 6 base cutter on one end and the 4 base cutter on the other end. In a further example, if an eight base cutter is combined with a four base cutter the complexity reduction is 128 fold. (See, Table 1 below). Similar calculations can be done for enzymes with different length recognition sequences.

[63] In yet another embodiment, further complexity reduction is achieved by designing at least one of the primers to hybridize at its 3' end to at least one nucleotide of the variable region of the restriction fragments to be amplified. The

additional selective nucleotide(s) further reduces the complexity of the resulting DNA pool because only those fragments which contain the complement of the specific nucleotide(s) incorporated in the PCR primer will be amplified. The level of complexity in the isolated pool can be varied depending upon the identity and number of nucleotides incorporated into the PCR primers and the restriction enzymes used to fragment the DNA. For example, if an enzyme with a 6 base recognition sequence is combined with an enzyme with a four base pair recognition sequence and only fragments that were digested with both enzymes are amplified on average 1 of every 8 restriction fragments will be amplified. This results in an 8 fold reduction in the complexity of the sample, reducing the 3,300 Mb human genome to 375 Mb. If an additional selective nucleotide is added to one of the primers only 1/4<sup>th</sup> of those fragments will be amplified or 1 out of every 32 restriction fragments, resulting in a 32 fold reduction in complexity. See, Table 1 and Table 2. Different combinations of selective nucleotides can be used to amplify different populations of SNPs. Table 2 illustrates the use of different selective nucleotides in combination with digestion with enzymes with a 6 and 4 base pair recognition sequence. A number of embodiments of the present invention involve amplification by PCR. Any of these embodiments may be further modified to reduce complexity using the above disclosed technique.

Table 1.

# of Selective Nucleotides	Number of bases in RE1 recognition site	Number of bases in RE2 recognition site	Fraction of RE1-RE2 Fragments	Complexity Reduction from Total Genome (Mb)
0	6	4	1/8	375
1	6	4	1/32	96
2	6	4	1/128	24
3	6	4	1/512	6
0	8	4	1/128	24
1	8	4	1/512	6

Table 2.

Selective nucleotides	Estimated Complexity	# of TSC SNPs Selected
O A	96Mb	24,000
A A	24Mb	6,000

A T	24Mb	6,000
T A	24Mb	6,000
C T	24Mb	6,000
A AT	6Mb	1,000
A AG	6Mb	1,000

[64] There are many known methods of amplifying nucleic acid sequences including *e.g.*, PCR. *See, e.g., PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, *et al.*, Academic Press, San Diego, CA, 1990); Mattila *et al.*, *Nucleic Acids Res.* 19, 4967 (1991); Eckert *et al.*, *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson *et al.*, IRL Press, Oxford); and U.S. Patent 4,683,202, 4,683,195, 4,800,159 4,965,188 and 5,333,675 each of which is incorporated herein by reference in their entireties for all purposes.

[65] PCR is an extremely powerful technique for amplifying specific polynucleotide sequences, including genomic DNA, single-stranded cDNA, and mRNA among others. Various methods of conducting PCR amplification and primer design and construction for PCR amplification will be known to those of skill in the art. Generally, in PCR a double stranded DNA to be amplified is denatured by heating the sample. New DNA synthesis is then primed by hybridizing primers to the target sequence in the presence of DNA polymerase and excess dNTPs. In subsequent cycles, the primers hybridize to the newly synthesized DNA to produce discreet products with the primer sequences at either end. The products accumulate exponentially with each successive round of amplification. Standard protocols may be found in, for example Sambrook *et al.* which is hereby incorporated by reference for all purposes.

[66] The DNA polymerase used in PCR is often a thermostable polymerase. This allows the enzyme to continue functioning after repeated cycles of heating necessary to denature the double stranded DNA. Polymerases that are useful for PCR include, for example, *Taq* DNA polymerase, *Tth* DNA polymerase, *Tfi* DNA polymerase, *Tma* DNA polymerase, *Tli* DNA polymerase, and *Pfu* DNA polymerase. There are many commercially available modified forms of these enzymes including: AmpliTaq® and AmpliTaq Gold® both available from Applied Biosystems. Many are available with

or without a 3- to 5' proofreading exonuclease activity. See, for example, Vent® and Vent® (exo-) available from New England Biolabs.

[67] Other suitable amplification methods include the ligase chain reaction (LCR) (e.g., Wu and Wallace, *Genomics* 4, 560 (1989) and Landegren *et al.*, *Science* 241, 1077 (1988)), transcription amplification (Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli *et al.*, *Proc. Nat. Acad. Sci. USA*, 87, 1874 (1990)) and nucleic acid based sequence amplification (NABSA). (See, US patents nos. 5,409,818, 5,554,517, and 6,063,603). The latter two amplification methods include isothermal reactions based on isothermal transcription, which produce both single-stranded RNA (ssRNA) and double-stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

[68] As those of skill in the art will appreciate, after isolation and amplification, the resulting sequences may be further analyzed using any known method including sequencing, HPLC, hybridization analysis, etc.

[69] Any of the above embodiments can be combined with *in silico* digestion techniques to predict the number and identities of SNPs predicted to be represented in a PCR product amplified using the above methods. For example, if restriction fragments that have been cleaved by *EcoRI* at one end and *Sau3A* at the other end are the subset to be amplified, a computer can be used to model an *EcoRI* and *Sau3A* digest of the genome to identify that subset of fragments and to identify the known or suspected SNPs that are present on that subset of fragments.

[70] The invention also includes kits for performing one or more of the different methods for analyzing a polynucleotide population described herein. Kits generally contain reagents necessary to perform the subject methods. The reagents may be supplied in pre-measured amounts for individual assays so as to increase reproducibility. Reagents may also be provided in a concentrated form.

[71] In one embodiment, the subject kits comprise adaptors and primers to amplify representative restriction fragments. The kits of the invention may also include one or more additional reagents required for various embodiments of the subject methods. Such additional reagents include, but are not limited to: restriction enzymes, DNA polymerases, DNA ligases, buffers, nucleotides and the like.

### METHODS OF USE

[72] The methods of the presently claimed invention can be used for a wide variety of applications. Any analysis of genomic DNA may be benefited by a reproducible method of complexity management. Furthermore, the methods and enriched fragments of the presently claimed invention are particularly well suited for study and characterization of extremely large regions of genomic DNA.

[73] In a preferred embodiment, the methods of the presently claimed invention are used for SNP discovery and to genotype individuals. For example, any of the procedures described above, alone or in combination, could be used to isolate the SNPs present in one or more specific regions of genomic DNA. Selection probes could be designed and manufactured to be used in combination with the methods of the invention to amplify only those fragments containing regions of interest, for example a region known to contain a SNP. Arrays could be designed and manufactured on a large scale basis to interrogate only those fragments containing the regions of interest. Thereafter, a sample from one or more individuals would be obtained and prepared using the same techniques which were used to prepare the selection probes or to design the array. Each sample can then be hybridized to an array and the hybridization pattern can be analyzed to determine the genotype of each individual or a population of individuals. Methods of use for polymorphisms and SNP discovery can be found in, for example, co-pending US application Nos. 08/813,159 and 09/428,350 which are herein incorporated by reference in their entirety for all purposes).

#### Correlation of Polymorphisms with Phenotypic Traits

[74] Most human sequence variation is attributable to or correlated with SNPs, with the rest attributable to insertions or deletions of one or more bases, repeat length polymorphisms and rearrangements. On average, SNPs occur every 1,000-2,000 bases when two human chromosomes are compared. (See, The International SNP Map Working Group, *Science* 409: 928-933 (2001) incorporated herein by reference in its entirety for all purposes.) Human diversity is limited not only by the number of SNPs occurring in the genome but further by the observation that specific combinations of alleles are found at closely linked sites.

[75] Correlation of individual polymorphisms or groups of polymorphisms with phenotypic characteristics is a valuable tool in the effort to identify DNA variation that contributes to population variation in phenotypic traits. Phenotypic traits include physical characteristics, risk for disease, and response to the environment. Polymorphisms that correlate with disease are particularly interesting because they represent mechanisms to accurately diagnose disease and targets for drug treatment. Hundreds of human diseases have already been correlated with individual polymorphisms but there are many diseases that are known to have an, as yet unidentified, genetic component and many diseases for which a component is or may be genetic.

[76] Many diseases may correlate with multiple genetic changes making identification of the polymorphisms associated with a given disease more difficult. One approach to overcome this difficulty is to systematically explore the limited set of common gene variants for association with disease.

[77] To identify correlation between one or more alleles and one or more phenotypic traits, individuals are tested for the presence or absence of polymorphic markers or marker sets and for the phenotypic trait or traits of interest. The presence or absence of a set of polymorphisms is compared for individuals who exhibit a particular trait and individuals who exhibit lack of the particular trait to determine if the presence or absence of a particular allele is associated with the trait of interest. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As an example of a correlation between a phenotypic trait and more than one polymorphism, it might be found that allele A1 at polymorphism A and allele B1 at polymorphism B correlate with a phenotypic trait of interest.

#### Diagnosis of Disease and Predisposition to Disease

[78] Markers or groups of markers that correlate with the symptoms or occurrence of disease can be used to diagnose disease or predisposition to disease without regard to phenotypic manifestation. To diagnose disease or predisposition to disease, individuals are tested for the presence or absence of polymorphic markers or marker sets that correlate with one or more diseases. If, for example, the presence of allele A1 at polymorphism A correlates with coronary artery disease then individuals with allele A1 at polymorphism A may be at an increased risk for the condition.

[79] Individuals can be tested before symptoms of the disease develop. Infants, for example, can be tested for genetic diseases such as phenylketonuria at birth.

Individuals of any age could be tested to determine risk profiles for the occurrence of future disease. Often early diagnosis can lead to more effective treatment and prevention of disease through dietary, behavior or pharmaceutical interventions.

Individuals can also be tested to determine carrier status for genetic disorders.

Potential parents can use this information to make family planning decisions.

[80] Individuals who develop symptoms of disease that are consistent with more than one diagnosis can be tested to make a more accurate diagnosis. If, for example, symptom S is consistent with diseases X, Y or Z but allele A1 at polymorphism A correlates with disease X but not with diseases Y or Z an individual with symptom S is tested for the presence or absence of allele A1 at polymorphism A. Presence of allele A1 at polymorphism A is consistent with a diagnosis of disease X. Genetic expression information discovered through the use of arrays has been used to determine the specific type of cancer a particular patient has. (See, Golub *et al.* *Science* 286: 531-537 (2001) hereby incorporated by reference in its entirety for all purposes.)

#### Pharmacogenomics

[81] Pharmacogenomics refers to the study of how your genes affect your response to drugs. There is great heterogeneity in the way individuals respond to medications, in terms of both host toxicity and treatment efficacy. There are many causes of this variability, including: severity of the disease being treated; drug interactions; and the individuals age and nutritional status. Despite the importance of these clinical variables, inherited differences in the form of genetic polymorphisms can have an even greater influence on the efficacy and toxicity of medications. Genetic polymorphisms in drug-metabolizing enzymes, transporters, receptors, and other drug targets have been linked to interindividual differences in the efficacy and toxicity of many medications. (See, Evans and Relling, *Science* 286: 487-491 (2001) which is herein incorporated by reference for all purposes).

[82] An individual patient has an inherited ability to metabolize, eliminate and respond to specific drugs. Correlation of polymorphisms with pharmacogenomic traits identifies those polymorphisms that impact drug toxicity and treatment efficacy. This information can be used by doctors to determine what course of medicine is best

for a particular patient and by pharmaceutical companies to develop new drugs that target a particular disease or particular individuals within the population, while decreasing the likelihood of adverse affects. Drugs can be targeted to groups of individuals who carry a specific allele or group of alleles. For example, individuals who carry allele A1 at polymorphism A may respond best to medication X while individuals who carry allele A2 respond best to medication Y. A trait may be the result of a single polymorphism but will often be determined by the interplay of several genes.

[83] In addition some drugs that are highly effective for a large percentage of the population, prove dangerous or even lethal for a very small percentage of the population. These drugs typically are not available to anyone. Pharmacogenomics can be used to correlate a specific genotype with an adverse drug response. If pharmaceutical companies and physicians can accurately identify those patients who would suffer adverse responses to a particular drug, the drug can be made available on a limited basis to those who would benefit from the drug.

[84] Similarly, some medications may be highly effective for only a very small percentage of the population while proving only slightly effective or even ineffective to a large percentage of patients. Pharmacogenomics allows pharmaceutical companies to predict which patients would be the ideal candidate for a particular drug, thereby dramatically reducing failure rates and providing greater incentive to companies to continue to conduct research into those drugs.

#### Determination of Relatedness

[85] There are many circumstances where relatedness between individuals is the subject of genotype analysis and the present invention can be applied to these procedures.

Paternity testing is commonly used to establish a biological relationship between a child and the putative father of that child. Genetic material from the child can be analyzed for occurrence of polymorphisms and compared to a similar analysis of the putative father's genetic material. Determination of relatedness is not limited to the relationship between father and child but can also be done to determine the relatedness between mother and child, (see *e.g.* Staub *et al.*, U.S. Pat. No.6,187,540) or more broadly, to determine how related one individual is to another, for example,



between races or species or between individuals from geographically separated populations, (see for example H. Kaessmann, *et al. Nature Genet.* 22, 78 (1999)).

#### Forensics

[86] The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (*e.g.*, by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance. A similar comparison of markers can be used to identify an individual's remains. For example the U.S. armed forces collect and archive a tissue sample for each service member. If unidentified human remains are suspected to be those of an individual a sample from the remains can be analyzed for markers and compared to the markers present in the tissue sample initially collected from that individual.

#### Marker Assisted Breeding

[87] Genetic markers can assist breeders in the understanding, selecting and managing of the genetic complexity of animals and plants. Agriculture industry, for example, has a great deal of incentive to try to produce crops with desirable traits (high yield, disease resistance, taste, smell, color, texture, etc.) as consumer demand increases and expectations change. However, many traits, even when the molecular mechanisms are known, are too difficult or costly to monitor during production. Readily detectable polymorphisms which are in close physical proximity to the desired genes can be used as a proxy to determine whether the desired trait is present or not in a particular organism. This provides for an efficient screening tool which can accelerate the selective breeding process.

## EXAMPLE

### *Step 1. Restriction enzyme digestion of genomic DNA*

[88] 2 µg human genomic DNA was digested with 20 units of *EcoR* I (New England Biolabs) and 20 units of *Sau3A* I in 80 µl 1x NEB buffer 4 with 1x BSA (NEB) by 37°C overnight incubation. The enzymes were inactivated at 65°C for 20 minute.

### *Step 2. Adaptor ligation*

[89] 500 ng genomic DNA digested in step 1 was ligated to 30 pmol *EcoR* I adaptors (5'-AGTATAAGGGGTTCG-3' (SEQ ID NO 1) and 5'-phosphate-AATTACGAACCCCTTATACT-3' (SEQ ID NO 2)), which leaves a 1 nucleotide gap between the 3' adaptor and the restriction fragment, and 300 pmol *Sau3A* I adaptors (5'-ATCGCCTTATAGTGAGTCG-3' (SEQ ID NO 3) and 5'-CGACTCACTATAAGGC-3' (SEQ ID NO 4)), which leaves a 1 nucleotide gap between the 5' end of the adaptor and the restriction fragment, with 4000 units of T4 Ligase (NEB) in 1x T4 Ligase buffer (NEB) at 16°C for overnight. The ligase was inactivated by incubation at 65°C for 10 minutes.

[90] Because the adaptors were designed to leave a gap in one strand (see Figure 3), only fragments that were cleaved by both enzymes will ligate both the *EcoR* I and *Sau3A* I adaptor sequences to the same strand and be amplified in the following step.

### *Step 3. Amplification with PCR*

[91] 4 ng genomic DNA ligated in step 2 was amplified with 2 units AmpliTaq Gold polymerase (PE) and 3 mM *EcoR* I primer (5'-AGTATAAGGGGTTCGTAATTCA-3' (SEQ ID NO 5)) and 3 mM *Sau3A* I primer (5'-ACTCACTATAAGGCGATCA-3' (SEQ ID NO 6)) in 50 µl 1x PCR buffer II (PE) with 2.5 mM MgCl<sub>2</sub>. PCR was performed in a MJ Research thermal cycler (Watertown, MA) with 35 cycles of 94°C for 30 seconds, 48°C for 30 seconds and 72°C for 2 minutes. 400 µl PCR reactions were pooled and purified and quantified by measuring their UV absorbance at 260 nm.

### *Step 4. Fragmentation, biotin labeling and hybridization to arrays*

[92] 5 µg purified PCR product was digested in 40 µl 1x One-Phor-All buffer (Pharmacia) with 0.6 unit DNase I (Promega) at 37°C for 15 minutes. The DNase I was inactivated by boiling for 15 minutes. 2 µl biotin-N6-ddATP (1.0 mM, NEN)

and 3 ul rTDT (15U/ $\mu$ l, GIBCO) were added to the DNA solution. The DNA fragments were labeled with biotin by incubation at 37°C for 1.5 hrs. Boiling for 15 minutes inactivated the enzyme. 20  $\mu$ g DNA was pooled, concentrated, mixed with hybridization cocktail and hybridized to arrays on a rotisserie at 49°C overnight.

*Step 5. Array washing, staining and scanning*

[93] Arrays were washed with 6X SSPE buffer on a fluidics station (Affymetrix) and 0.1x MES buffer (Affymetrix) at 45°C for 30 minutes, stained and scanned under standard procedures.

CONCLUSION

[94] From the foregoing it can be seen that the present invention provides a flexible and scalable method for analyzing complex samples of DNA, such as genomic DNA. These methods are not limited to any particular type of nucleic acid sample: plant, bacterial, animal (including human) total genome DNA, RNA, cDNA and the like may be analyzed using some or all of the methods disclosed in this invention. This invention provides a powerful tool for analysis of complex nucleic acid samples. From experiment design to isolation of desired fragments and hybridization to an appropriate array, the above invention provides for fast, efficient and inexpensive methods of complex nucleic acid analysis.

[95] All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.

We claim:

1. A method of reducing the complexity of a nucleic acid sample comprising:  
fragmenting the nucleic acid sample using a first and second restriction enzyme to produce fragments;  
ligating adaptors to the fragments; and  
selectively amplifying the fragments that were cut on one end by the first restriction enzyme and on the other end by the second restriction enzyme.
2. The method of claim 1 wherein the amplified fragments comprise at least 0.01% of the first nucleic acid sample.
3. The method of claim 1 wherein the amplified fragments comprise at least 0.5% of the first nucleic acid sample.
4. The method of claim 1 wherein the amplified fragments comprise at least 3% of the first nucleic acid sample.
5. The method of claim 1 wherein the amplified fragments comprise at least 12% of the first nucleic acid sample.
6. The method of claim 1 wherein the amplified fragments comprise at least 30% of the first nucleic acid sample.
7. The method of claim 1 wherein the amplified fragments comprise at least 50% of the first nucleic acid sample.
8. The method of claim 1 wherein the first restriction enzyme has a six base pair recognition sequence and the second restriction enzyme has a four base pair recognition sequence.
9. The method of claim 1 wherein the first restriction enzyme has an eight base pair recognition sequence and the second restriction enzyme has a four base pair recognition sequence.

10. The method of claim 1 wherein the fragments are amplified by PCR.
11. The method of claim 1 wherein the nucleic acid sample is genomic DNA, DNA, cDNA derived from RNA, total RNA or mRNA.
12. The method of claim 1 wherein ligation of one strand of each adaptor is blocked.
13. The method of claim 12 wherein ligation is blocked by introducing a gap of at least one nucleotide between one strand of the adaptor and one strand of the fragment.
14. The method of claim 12 wherein ligation is blocked by the absence of a phosphate at the 5' end of an adaptor strand.
15. The method of claim 12 wherein ligation is blocked by the presence of a modified nucleotide at the 5' or 3' end of an adaptor strand.
16. The method of claim 12 wherein ligation is blocked by a terminal modification in one strand of an adaptor.
17. The method of claim 12 wherein ligation is blocked at the 5' end of one strand of one adaptor and at the 3' end of one strand of the other adaptor.
18. The method of claim 12 wherein ligation is blocked at the 5' end of one adaptor and at the 3' end of the other adaptor.
19. The method of claim 1 wherein one adaptor comprises a 5' overhang comprising a primer binding site and the other adaptor comprises a 3' overhang comprising a primer binding site.
20. A method for analyzing a nucleic acid sample comprising:  
fragmenting the nucleic acid sample using a first and second restriction enzyme to produce fragments;

ligating adaptors to the fragments;  
selectively amplifying the fragments that were cut on one end by the first restriction enzyme and on the other end by the second restriction enzyme;  
providing a nucleic acid array;  
hybridizing the amplified fragments to the array; and  
analyzing a hybridization pattern resulting from the hybridization.

21. The method of claim 20 wherein the method for analyzing the nucleic acid sample comprises determining whether the nucleic acid sample contains sequence variations.
22. The method of claim 21 wherein the sequence variations are single nucleotide polymorphisms (SNPs).
23. The method of claim 20 wherein the nucleic acid array is designed to query DNA fragments which have been produced by the procedures used to obtain the amplified fragments.
24. The method of claim 20 wherein a substantial amount of the sequences predicted to be contained in the amplified fragments are first determined by a computer system.
25. A method of screening for DNA sequence variations in an individual comprising:
  - providing a nucleic acid sample from the individual;
  - fragmenting the nucleic acid sample using a first and second restriction enzyme to produce fragments;
  - ligating adaptors to the fragments; and
  - selectively amplifying the fragments that were cut on one end by the first restriction enzyme and on the other end by the second restriction enzyme;
  - providing a nucleic acid array;
  - hybridizing the amplified fragments to the array;
  - generating a hybridization pattern resulting from the hybridization; and

determining the presence or absence of DNA sequence variations in the individual based upon an analysis of the hybridization pattern.

26. The method of claim 25 wherein the sequence variation is at least one single nucleotide polymorphism (SNP).

27. The method of claim 26 wherein the at least one SNP is associated with a disease.

28. The method of claim 26 wherein the at least one SNP is associated with the efficacy of a drug.

29. A method of screening for DNA sequence variation in a population of individuals comprising:

providing a first nucleic acid sample from each of the individuals;

providing a second nucleic acid sample by:

fragmenting the first nucleic acid sample using a first and second restriction enzyme to produce fragments;

ligating adaptors to the fragments; and

selectively amplifying the fragments that were cut on one end by the first restriction enzyme and on the other end by the second restriction enzyme;

providing a plurality of identical nucleic acid arrays wherein the arrays comprise probes designed to interrogate for DNA sequence variations;

hybridizing each of the second nucleic acid samples to one of the plurality of identical arrays;

generating a plurality of hybridization patterns resulting from the hybridizations; and

analyzing the hybridization patterns to determine the presence or absence of sequence variation in the population of individuals.

30. The method of claim 29 wherein the sequence variation is at least one single nucleotide polymorphism (SNP).

31. The method of claim 27 wherein the at least one SNP is associated with a disease.
32. The method of claim 27 wherein the at least one SNP is associated with the efficacy of a drug.
33. A method of genotyping an individual comprising:  
providing a first nucleic acid sample from the individual;  
providing a second nucleic acid sample by:  
    fragmenting the first nucleic acid sample using a first and second restriction enzyme to produce fragments;  
    ligating adaptors to the fragments; and  
    selectively amplifying the fragments that were cut on one end by the first restriction enzyme and on the other end by the second restriction enzyme;  
hybridizing the second nucleic acid sample to an array designed to determine the presence or absence of one or more alleles of a collection of SNPs;  
generating a hybridization pattern resulting from the hybridization; and  
determining the presence or absence of the one or more alleles of the collection of SNPs.
34. A kit for reducing the complexity of a nucleic acid sample comprising:  
buffers and restriction enzymes for fragmenting a nucleic acid sample,  
a ligase and adaptors to be ligated to the fragments, the adaptors being designed for the selective amplification of the fragments that were cut on one end by a first restriction enzyme and on the other end by a second restriction enzyme,  
primers for the selective amplification, and  
instructions for the use of the kit.



1/4

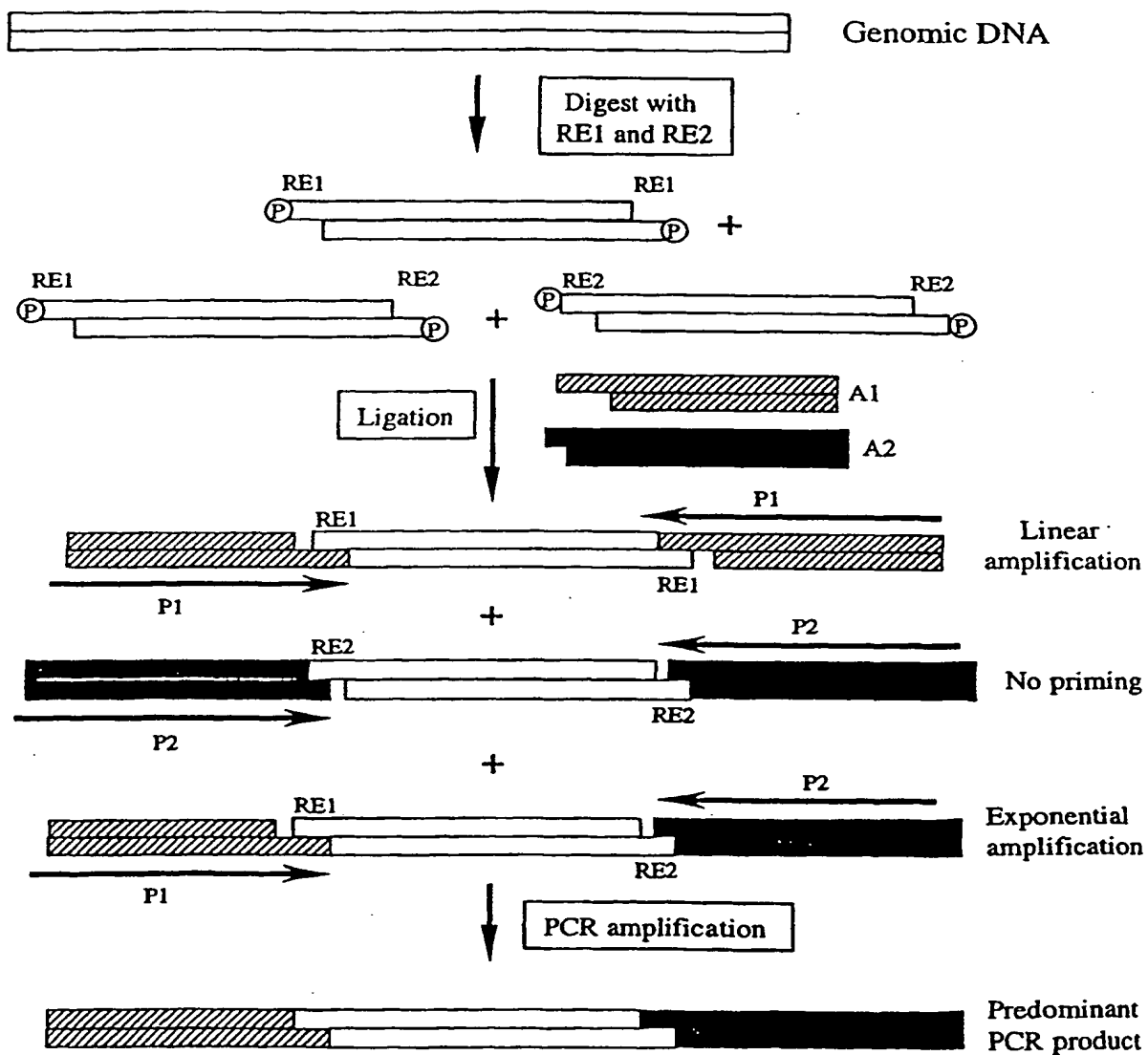
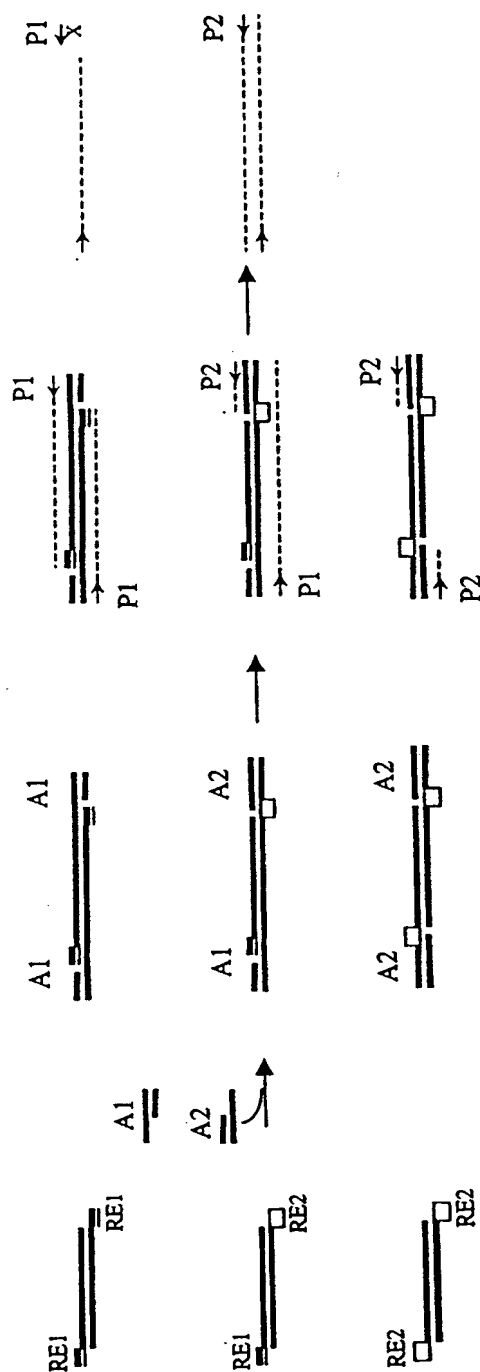


Figure 1

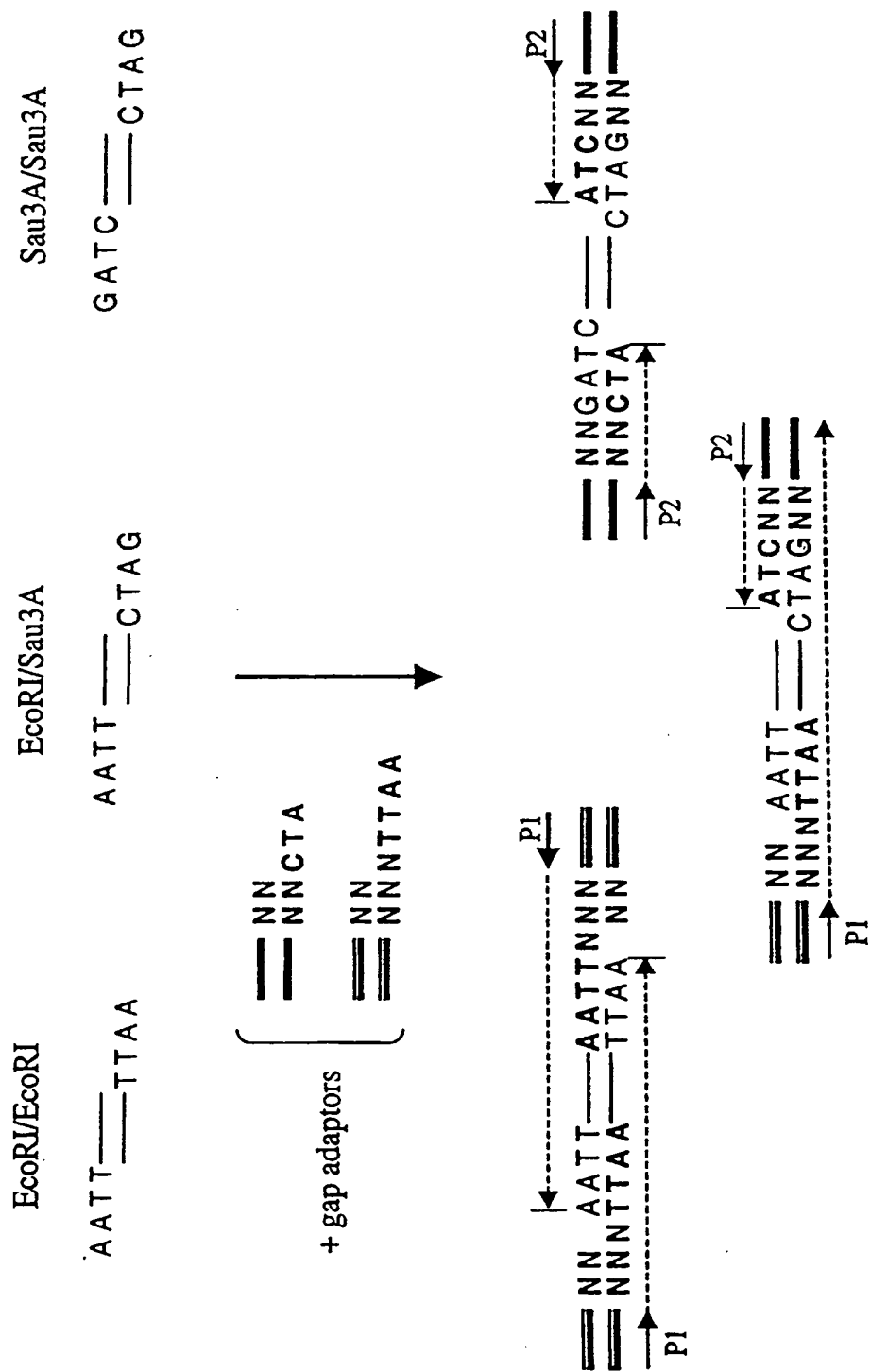
2/4

Figure 2



3/4

Figure 3



4/4

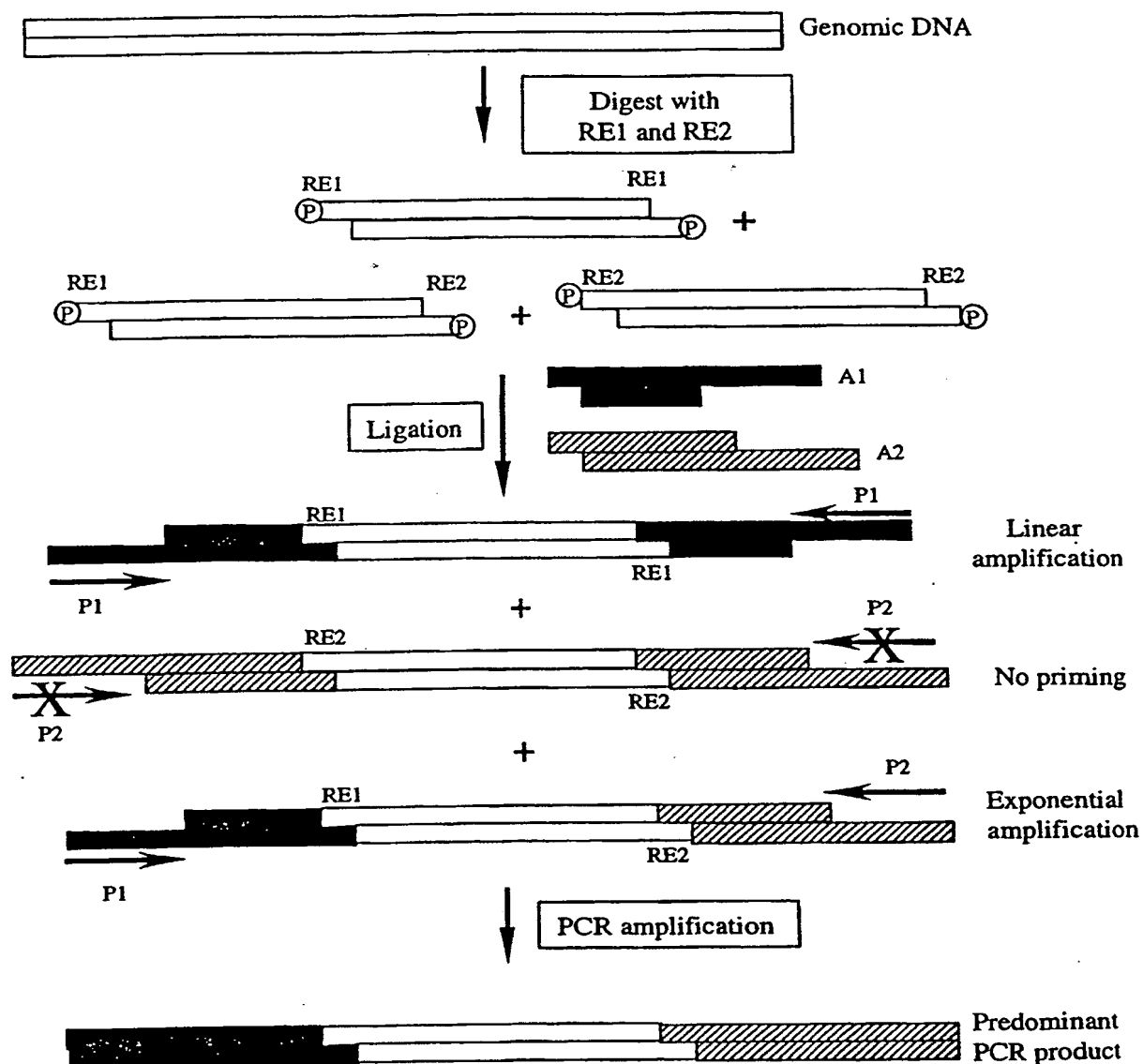


Figure 4

# INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/23568

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : C12P 19/34

US CL : 435/91.1

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/91.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)  
Please See Continuation Sheet

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SMITH. D.R. Ligation-mediated PCR of restriction fragments from large DNA molecules. PCR Methods and Applications. 1992. Vol 2. No 1, pages 21-27., see entire document.	1-7, 10-11
---		-----
Y		8-9, 20-34
Y	US 6,156,502 A (BEATTIE et al) 05 December 2000 (5.12.2000), column 3, lines 13-28, see entire document.	20-34

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

\* Special categories of cited documents:

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier application or patent published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- "&" document member of the same patent family

Date of the actual completion of the international search

19 September 2002 (19.09.2002)

Date of mailing of the international search report

03 OCT 2002

Name and mailing address of the ISA/US

Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703)305-3230

Authorized officer

Valerie Bell-Harris for  
Gary Jones

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)

# INTERNATIONAL SEARCH REPORT

PCT/US02/23568

## Continuation of B. FIELDS SEARCHED Item 3:

databases: MEDLINE, BIOSIS, CAPLUS, US PATENT DATABASE

search terms: DNA, amplification, amplify, amplified, ligate, ligated, ligation, adapter, adaptor, specific, selective, digest, digestion, restriction, restrict, endonuclease, block, blocked, extend, extension, polymerization, prevent, prevention